「ゲノム情報解析基礎」4/23 分**課題とその回答**、および**感想やコメントなど**の一部に対するコメント

4 月 24 日、門田幸二

1. 課題とその回答

課題 1：(contig_1, contig_2, contig_3 の配列長がそれぞれ 24 bp, 103 bp, 65 bp だったときの N50 の値を示せ)
正答：(103+65+24) / 2 = 96bp。長いものから並べていったときに 96bp に達したときのコンティグの長さが N50 なので、**N50 = 103 bp**。

課題 2：(「**(R で)塩基配列解析**」のウェブページの「**イントロダクション| 一般 | multi-fasta ファイルから指定した配列長のもののみ抽出**」の 1.の下記スクリプトを改変して、param で**指定した配列長未満のものを抽出したい**場合にどこをどう変更すればよいか示せ)

```
------    ここから    ------
in_f <- "hoge4.fa"
out_f <- "hoge1.fasta"
param <- 50

library(Biostrings)
reads <- read.DNAStringSet(in_f, format="fasta")
reads
reads <- reads[width(reads) >= param]
reads
write.XStringSet(reads, file=out_f, format="fasta")

------    ここまで    ------
```

正答：下から三行目の記述を以下のように変更。
`reads <- reads[width(reads) < param]`
「未満」と書いてあるので、"<"とすべきです。"<="だったり"=<"などとかいていた人は減点です。これは「未満」ではなく「以下」ですので…。

課題 3：vmatchPattern 関数はミスマッチを許容するオプションが存在するか？（回答は基本的に「yes or no」でよいが、利用可能なオプションを実際に試すなどしてみてください）
正答：存在する（つまり yes）。配布資料のスライド 60 からも読み取れますが、min.mismatch や max.mismatch というオプションが存在します。これらのオプションのデフォルトは min.mismatch=0 や max.mismatch=0 となっているので、ここを明示的に指定しなければ 100%一致のもののみが得られることになりますが、1 とか 2 の値を指定することでミスマッチ数を許容したパターンマッチングを行うことができます。

2. 感想やコメントなどの一部に対するコメント

基本的に好意的なコメントを多くいただき、講義準備を入念に行ったかいがありました。ありがとうございます。ちょっとレスしとこうと思った方のみ、そのコメントのエッセンスのみ掲載し、そのレス、という形式にしています。受講生 ID 順にしてあります。下三桁のみグループ化して記載してあります。基本的に走り書きですので、細かい言い回しなどについては好意的に解釈していただければ幸いです m(_ _)m。

001-050：ゆっくりでわかりやすかったがもう少し詳しく説明してほしかった。
レス：9 月にある農学生命科学特論 I のほうでも一部復習を兼ねて説明する予定です。が講義の中での説明は限界がありますので、不明な点があればメールなどで直接問い合わせてください。

001-050：課題もウェブにアップして。後で確認できるから
レス：というわけでアップしました。

001-050：HP のここから、ここまでの先頭にも#を入れると便利ではないか？
レス：一理あります。昔そうしようと思ったこともありました。しかし、一つの項目の中で例題が複数ある場合などに、初心者の方が複数個まとめてコピーしようとしたこともあり、#をいれるとそういうケアレスミスに気づくのがより遅れます。なので、結局やめました。

001-050：手を動かすところが多くてよかったが、関数の説明をもう少しして。
レス：講義の中での説明は限界がありますので、このあたりについては、基本的に「?関数名」でマニュアルを自分で読み込むことで理解を深める、ようにしてください。もちろん、それでも不明な点があれば直接問い合わせてください。

051-100：使用するファイルは先にダウンロードさせておいたほうがよりスムーズかもね。
レス：一理あります。が、4/23 も 105 名の出席者がおり、それだけの大人数だとサーバーに負担がかかって、講義の最初の段階でネットワーク自体がだめになる可能性を懸念したことと、「(R で）塩基配列解析」のウェブページから、自分でサンプルファイルのダウンロードからやってもらう、という手順を教えたかったためです。

051-100：Excel, Word, メモ帳のショートカットをデスクトップ上においてくれるとやりやすい
レス：おっしゃる通り。少なくとも年度初めの段階ではそうなるように善処させていただきます。

101-150：N50 がなぜ信頼度の指標になるかなどをもう少し詳しく説明してもらえるとよかった。
レス：なぜ信頼度の指標になるかは、なかなか答えづらいですね。たぶん、だれかこの分野の大御所が提唱して、Average length や Median length の他に N50 を一つくらい追加しても、、、まあいいんじゃない、、、的なノリだったのかもしれません。。。嶋田先生の講義でもありましたが、N10, N20, …, N90 など他にもいろいろありますので、one of them 的な理解でいいと思います。データの解釈の仕方、については人それぞれなので、なかなか講義の枠組みではしゃべりにくい、、、です。

101-150：sequence logos の計算式の妥当性が不明
レス：あの数式で表現するのが sequence logos です。ぶっちゃけ、私もあの数式で全てが表現できているとは思ってないです。よって、この改良版の数式を提案する予定です。これがバイオインフォマティクスです。

101-150：R studio は使わないの？
レス：そういうものがある、ということは知っていますが R で特に不便を感じていないので、、、。

151-200：前回の講義に出席できなかったんだけど。。。
レス：前回の講義資料を事務局までとりにくるなりしてくださいね。

151-200：延長コードがほしい。。。
レス：コードに足を引っ掛けて大惨事、、、という事態を避けるため、二コマ連続の講義程度では PC のバッテリのへたれを実感しない限り、延長コードを用意する予定は残念ながらありません。。。